“惩前毖后”与“率先垂范”：第三方干预行为的影响效应[1]

郭禹辰　刘艳彬　程 远

浙大宁波理工学院商学院，宁波 315100

摘　要　第三方干预是维持和发展社会规范的关键力量，对整个人类群体具有进化意义。补偿和惩罚是第三方干预的两种主要形式，二者均是维护社会规范的重要力量，具有恢复得失平衡以及促进规范遵从的积极作用。研究者广泛认可威慑效应是第三方惩罚促进规范遵从的主要作用机制，然而也有许多研究结果与该假说相悖。除威慑作用外，第三方干预行为作为高代价信号，也具有澄清社会规范、改变人们规范知觉的作用，这暗示着信号效应可能也是第三方干预促进规范遵从的重要作用机制。探究第三方惩罚促进规范遵从的边界条件、检验第三方补偿等非破坏性措施在维护社会规范方面的有效性是未来研究的重要方向。

关键词　第三方干预，第三方惩罚，第三方补偿，社会规范，威慑效应，信号效应

分类号　B849：C91

收稿日期：2023-4-18

* 国家社会科学基金重大项目资助(21&ZD184)

通信作者：程远，E-mail：ycheng@zju.edu.cn

1 引言

《孟子·离娄》中有言：“不以规矩，不能成方圆”。在漫长的社会发展进程中，人类逐渐形成了对彼此行为方式的期望与承诺(Tomasello & Vaish, 2013)，这些达成共识的行为标准就是社会规范，其在促进人类合作进化、维持社会秩序方面起到决定性作用(Bernhard et al., 2006)。人们不仅承诺自己会遵守规范，也默认他人亦做出了遵守规范的承诺。因此在发现他人违背规范时，即使事不关己，甚至代价高昂，人们也会自发地进行干预来维护社会规范。研究者将这种自发行为称为第三方干预，并认为这种看似“非适应性”的行为是社会规范得以维持和发展的关键力量(e.g., Fabbri & Carbonara, 2017; Fehr & Fischbacher, 2004)，对整个人类群体具有重要的进化意义(Fehr & Williams, 2018)。

过去二十年来，第三方干预促进社会规范的研究取得了丰硕的成果，然而在理论层面上还缺少对第三方干预作用机制的系统梳理和总结。早期研究者普遍认为，第三方干预不仅通过提供惩罚或补偿来确保人际互动的结果符合社会规范要求，也依靠惩罚威慑阻止违背社会规范的行为再次发生(e.g., Fehr & Gächter, 2002; Robinson & Darley, 2003)。近年来，有研究者提出了不同观点，认为惩罚性干预作为外部激励虽然能够在短期内促进规范遵从，但是也会破坏人际信任、排挤人们遵守规范的内部动机，致使惩罚撤销后不良行为迅速反弹(Mulder et al., 2006; 陈思静 等, 2015)，无法真正起到促进社会规范的作用。此外，惩罚性干预的效果也取决于其正当性，当惩罚被滥用时（如实施反社会惩罚），该行为不仅无法有效维护规范，反而会损害正当惩罚的积极效果，导致群体合作水平下降(Herrmann et al., 2008; Fatas & Mateu, 2015)。反而，不具威慑力的恢复性干预可能在维护社会规范方面起到出乎意料的良好效果(Wiessner, 2020)。这些研究观点与结果的分歧源自何处？为了回答这一问题，有必要梳理第三方干预行为在维护社会规范方面的具体效果，并深入考察其内部作用机制。

本文基于威慑和信号的双重视角对第三方干预维护社会规范的功能和作用机制进行了系统回顾，并展望了未来的研究方向。

2 第三方干预的概念

“路见不平，拔刀相助”是人们耳熟能详的俚语，描绘了中国传统文化中锄强扶弱的侠义行为，体现了自古以来民间的朴素正义。这种在目睹负面事件发生后，自发伸张正义的行为在学界被称为第三方干预(Gummerum et al., 2016)，因其具有较强的利他属性，又被称为第三方利他行为(Gordon et al., 2014; Liu et al., 2018)。根据指向对象不同，第三方干预行为可以分为第三方惩罚(Third-party Punishment, TPP)和第三方补偿(Third-party Compensation, TPC)

两种主要形式，即利益无关的第三方愿意自己付出代价，以对造成伤害的违规者做出惩罚(Fehr & Gächter, 2002; Fehr & Fischbacher, 2004)，或对受到伤害的受害者进行补偿(Leliveld et al., 2012; Lotz et al., 2011)。第三方干预在生活中广泛存在，例如将肇事逃逸的车牌拍下递交给警方，或简单地口头谴责一个插队者，都是典型的第三方惩罚行为；而为暴力事件受害者提供无偿法律援助，或者将花圃中的饮料瓶捡起扔进垃圾桶，都属于第三方补偿行为。

第三方干预虽然是一种长期存在于人类社会中的亲社会行为(Schroeder et al., 2003; 徐杰 等, 2017)，但直到近二十年来才为学界所关注。Fehr 及其合作者们开发了基于博弈任务的第三方观察者范式，率先在实验室内对第三方惩罚行为进行了定量研究。他们在独裁者博弈实验中发现，即使被试自己不参与独裁者博弈，也会在观察到不公平的独裁者分配后，选择自己付出代价以减少独裁者的收益(Fehr & Gächter, 2002)。在随后的公共物品博弈实验中，他们再次发现了第三方惩罚的存在，并认为该行为可能是维持人类合作的关键力量(Fehr & Fischbacher, 2004)。此后，研究者对第三方惩罚这种违背“理性经济人”假说的行为展现出浓厚兴趣，就其行为动机、强度、方式、功能、影响因素以及神经基础等方面展开了大量研究(e.g., Bellucci et al., 2020; Ginther et al., 2016; Henrich et al., 2006; Jordan et al., 2016; 陈思静 等, 2015; 陈思静, 徐烨超, 2020; 刘映杰 等, 2022; 谢东杰, 苏彦捷, 2019)。

随着研究深入，研究者发现惩罚并不是恢复公正的唯一手段(Chavez & Bicchieri, 2013)，除了惩罚违规者，第三方观察者也关心受害者的福祉，愿意付出个人代价来弥补受害者的损失(e.g., Liu et al., 2018; Lotz et al., 2011; Thulin & Bicchieri, 2016; 徐杰 等, 2017; 王华根 等, 2020)，或者同时惩罚违规者以及补偿受害者(e.g., Van Doorn et al., 2014; Van Doorn et al., 2018)。尤其当第三方具备较高的特质性共情时，他们会更倾向于补偿受害者而非惩罚违规者(Hu et al., 2015; Leliveld et al., 2012)。

第三方可以通过多种方式实施干预。诸如对违规者进行口头谴责、传播流言、社会排斥、拒绝帮助、生理惩罚(如噪音、击打)或是金钱惩罚等都属于第三方惩罚行为(e.g., Balafoutas et al., 2016; Dimitroff et al., 2020; Fehr & Fischbacher, 2004; Feinberg et al., 2014; Jordan et al., 2014; Marshall et al., 2019; Marshall et al., 2021)；而为受害者提供帮助、提供信息支持或给予金钱补偿等则属于第三方补偿行为(Chavez & Bicchieri, 2013; Hart et al., 2019; Liu et al., 2018; Lotz et al., 2011; 王华根 等, 2020)。人们在现实生活中经常采用传播流言蜚语和社交回避的方式来进行第三方惩罚(Molho et al., 2020)，或是采取恢复性措施来进行第三方补偿(Wiessner, 2020)。但受限于现实中第三方干预行为的高度情境特异性，多数实证研究采用经济博弈范式，即以金钱惩罚或补偿作为第三方干预的执行方式(e.g., Kamei, 2018, 2020; Lewisch et al.,

2011; Stagnaro et al., 2017)。

3 第三方干预的社会规范维护功能

社会规范是人类社会和经济秩序的基础(Hume, 1888)，通过约定群体成员在特定情境下的行为准则，降低社会经济交换的交易成本、促进非亲缘个体间合作(Baumgartner et al., 2012; Henrich et al., 2006; Wiessner, 2020)。社会规范作为一种社会"隐性共识"，主要依赖社会成员的自我约束。反之，人们也有可能在利益驱动下做出违背社会规范的行为。在无法依赖于法律规章等正式系统制裁的情况下，社会规范之所以能够得以长期维持，离不开第三方干预的力量(Schroeder et al., 2003; Tomasello & Vaish, 2013)。无论是惩罚违规者，还是补偿受害者，都是社会成员对违背规范行为的回应，可以起到恢复社会公平，维护社会规范的作用(Leliveld et al., 2012; van Prooijen, 2010; Van Doorn et al., 2014)。总体而言，第三方干预的社会规范维护功能体现在两个方面上，一是恢复得失平衡，二是促进规范遵从。

3.1 恢复得失平衡功能

第三方干预可以通过恢复得失平衡来维护社会规范，即通过惩罚违规者或补偿受害者，使情况恢复到社会规范约定下的"应有"状态，这一功能主要与报应主义的行为动机相对应。恢复平衡包括使违规者承担其所应得的后果和报应(Carlsmith, 2006; Carlsmith et al., 2002)，以及使遭受损失的人或事物从被伤害的状态中恢复(Schroeder et al., 2003; Thulin & Bicchieri, 2016)。该动机是一种"应得"视角下的动机(Marshall et al., 2021)，即人们希望做了坏事的人得到应有的报应，因此会根据违规者造成的伤害给予他们相应的惩罚；同时，人们也希望无辜遭受伤害的人得到补偿，因此会根据受害者遭受损失的程度给予他们相应的补偿或帮助。研究发现，在不受限制的情况下，人们往往同时实施第三方惩罚以及第三方补偿来恢复正义(Lotz et al., 2011; Van Doorn et al., 2018)；而当只能选择一种干预方式时，人们整体上更倾向于对受害者做出补偿(Dhaliwal et al., 2021)，并且其实际行为决策经常受到自身人格特质(Leliveld et al., 2012; Hu et al., 2015)和得失情境框架(Liu et al., 2019)的影响。

由于干预的目标在于恢复平衡，因此第三方干预的力度经常以错误和伤害的程度为标准。研究发现，违规行为背离规范的程度越高、受害者遭受的损失越大，第三方做出惩罚或补偿的力度就越大(e.g., Fabbri & Carbonara, 2017; Heffner & FeldmanHall, 2019; Kamei, 2018; Lewisch et al., 2011; Ouyang et al., 2021)。例如，Rodrigues 等人(2020)的研究显示，独裁者的公平程度正向预测第三方惩罚与补偿的力度：当独裁者做出相对不公平的分配时，独裁者会被扣除约 21%的收益，接受者会被补偿约 35%的收益；而当独裁者做出非常不公平的分配

时，独裁者会被扣除约 29%的收益，接受者会被补偿约 44%的收益。此外，第三方经常基于规范的要求来调整干预力度，比如通过惩罚或补偿，将各方的不平等收益调整到满足公平规范(Chavez & Bicchieri, 2013)，或是以违规者的实际收益或受害者实际损失作为参照点来决定惩罚和补偿金额(Koenig & Riley, 2017)。

3.2 促进规范遵从功能

第三方干预也通过促使社会成员遵守规范、抑制规范违背行为来维护社会规范，这一功能主要与结果主义的行为动机相对应。结果主义行为动机是一种“预防”视角下的动机，主要与第三方惩罚行为有关，即人们做出惩罚是为了震慑潜在违规者，从而避免违规行为再次发生，因此它也被称为功利主义动机(Akers, 1990; Tan & Xiao, 2018)。许多研究支持了第三方惩罚的“功利主义”功能，发现第三方惩罚能有效提高人们在社会互动时的公平程度(Henrich et al., 2006; Martin et al., 2021)、合作程度(Fehr & Gächter, 2002; Stagnaro et al., 2017)以及互惠程度(Charness et al., 2008)。在对儿童的研究中也发现了一致的结果，如第三方惩罚可以显著提高儿童在囚徒博弈中的合作水平(Lergetporer et al., 2014)，以及儿童在多轮任务中选择公平分配的可能性(Martin et al., 2021)。

然而，也有研究发现第三方惩罚并不总是能够促进规范遵从，特别是在第三方惩罚未能持续存在，或者第三方惩罚缺乏正当性的时候，其维护规范的效果会明显减弱。首先，第三方惩罚在促进规范遵从方面的效果经常表现出“有益于当下，有损于未来”的特点，即人们迫于第三方惩罚的压力而遵守规范，却在惩罚撤销后立刻原形毕露，甚至变本加厉。这是因为惩罚作为一种外部激励，虽然能立即改变人们行为，却也会排挤人们遵守规范的内部动机。研究发现，人们虽然在有惩罚阶段表现出较高的合作水平，但在惩罚解除后其合作水平便迅速下降，甚至远低于未经历惩罚者(Rand et al., 2009; 陈思静 等, 2015)。有研究者认为，人际信任遭到破坏以及内部动机削弱可能是导致该现象的重要原因(Mulder et al., 2006; Xiao & Houser, 2011)。进化动力学研究也发现，虽然第三方惩罚可以促使发展中社会更快地向高度合作的社会过渡，但是当社会已经进入合作状态后，过度重视惩罚会导致更多社会损失(Yu et al., 2016)。此外，当第三方惩罚缺乏正当性时，也无法有效维护社会规范。无论在现实生活还是实验室中，第三方惩罚都存在被滥用的可能性，即惩罚指向了亲社会的合作者而非违规者。研究者在不同文化中都发现了这种反社会惩罚的存在(Herrmann et al., 2008)，并认为该行为是一种出于报复性策略(e.g., Sylwester et al., 2013; van Dijk et al., 2015)或竞争性策略(Basurto et al., 2016; Pleasant & Barclay, 2018)的负面行为。这种缺乏正当性的第三方惩罚不仅会降低人际信任、破坏群体合作、削弱群体利益(e.g., Fatas & Mateu, 2015; Herrmann et al.,

2008; Mulder et al., 2006)，也可能暗示群体中存在大量违规行为(陈思静, 朱玥, 2020)，继而诱发更多的违背社会规范行为。研究发现，只有在排除了反社会惩罚的可能性之后，惩罚才能有效提高合作水平(Rand & Nowak, 2011)。

第三方补偿在维护规范方面的作用同样存在争议。有研究者认为，只有第三方惩罚才能够威慑违规行为、稳定群体内合作(Balliet et al., 2011; Mathew & Boyd, 2011)，而第三方补偿没有降低违规行为的适应度（违背规范行为没有受到惩罚，仍是利益最大化的），因此并不能抑制违规行为再次发生(Chavez & Bicchieri, 2013)。然而也有研究者持有不同观点，认为第三方补偿作为惩罚的替代性干预措施，在维护规范方面的作用被忽视了(Almenberg et al., 2010)，特别是当违规行为不太严重时，补偿在维护社会规范方面的效果可能比惩罚更好(Wiessner, 2020; 冯佳兴, 2020)。比如，第三方补偿不仅能促进合作行为(Bottom et al., 2002)，也能促进信任的修复(De Cremer, 2010)。通过对恩加(恩加是新几内亚所属城市，存在着多元的司法体系，长久以来形成了一种依赖于群体的第三方补偿机制)当地 3 个村庄法庭在 10 年间的数百起案例进行定性分析，研究者发现相比实施惩罚，恩加人更愿意通过弥补已造成的伤害来维持公平，而这种恢复性措施对维持社会秩序和良好人际关系都具有重要作用(Wiessner , 2020)。

总之，基于不同的哲学视角，第三方干预在维护社会规范方面的影响效应可以具体分为恢复得失平衡和促进规范遵从两大功能(见图 1)，前者毋庸置疑，后者却尚且存在一定争议。为进一步解释第三方干预对规范遵从行为影响的矛盾性结果，有必要就其内部作用机制展开更深入的探讨。

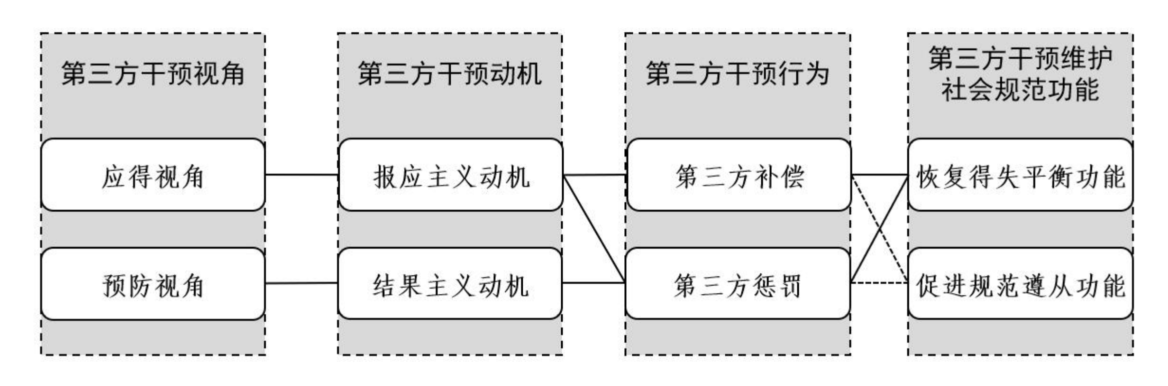


图 1 第三方干预的哲学视角与功能

4 第三方干预促进规范遵从的作用机制

4.1 第三方惩罚的威慑效应

威慑作为第三方惩罚的重要作用机制，似乎符合人们的直觉经验。当有人因为随地吐痰

遭到他人谴责时，可能会感到羞愧和尴尬。为避免再次体验到类似的不愉快情绪，他未来可能不会再随地吐痰。同时，第三方惩罚对违规行为的抑制作用也可能溢出到其他社会成员身上，比如当其他人看到有人因为随地吐痰而被谴责时，他们也不太可能会随地吐痰。

“威慑”是犯罪学领域的核心概念之一，被定义为“出于对法律惩罚的畏惧而不做出或少做出犯罪行为”(Gibbs, 1968)。威慑理论假定潜在罪犯是一个理性的行动者，会对犯罪的成本和收益进行理性计算，因此当其预期受到惩罚带来的痛苦和损失将抵消，甚至高过犯罪带来的收益时，犯罪就不会发生(Akers, 1990)。根据威慑对象的不同，威慑可以分为特殊威慑(Specific Deterrence)和一般威慑(General Deterrence)两种方式，前者是指对那些直接受到惩罚者的威慑效应，后者则是指对其他社会成员(即潜在罪犯)的威慑效应。也就是说，惩罚可以从两方面阻止违法行为：首先，对违规者的惩罚会使得他们产生对未来类似惩罚的恐惧，从而阻止重复犯罪；其次，对违规者施加的惩罚会导致其他人害怕类似的惩罚，从而阻止其他人犯下类似的错误(Robinson & Darley, 2003)。

社会学习理论也阐释了与一般威慑类似的观点。班杜拉指出：“人们可以从他人的成功或错误经验中学习，观察到的结果可以像直接经历的结果一样改变人们自己的行为。”换句话说，当个体观察到他人因做出某些行为而受到惩罚时，会产生自己做出类似行为也会招致惩罚的预期，进而避免做出类似行为，这个过程被称为替代性惩罚(Bandura, 1986)。元分析研究发现，替代性惩罚的效果显著：观察到他人因某种行为受到惩罚的人们将更少做出类似的行为(Malouff et al., 2009)。

根据威慑理论和替代性惩罚的观点，第三方惩罚应当能够强化社会规范。一方面，第三方惩罚震慑了违规者，提醒他们在未来遵守规范以免再次遭到惩罚。另一方面，第三方惩罚也暗示其他社会成员，当前行为是不被允许的，不遵守规范将付出代价(Delton & Krasnow, 2017)。出于对预期惩罚的畏惧，其他成员也会避免做出违背规范的行为。有研究者指出，威慑是第三方惩罚的主要动机之一(Carlsmith, 2006; Fehr & Gächter, 2002; Hauert et al., 2007)。人们之所以宁愿牺牲个人利益也要惩罚违规者，部分原因是出于功利视角的考量，也就是威慑潜在违规者以敦促其改正行为，从而降低潜在违规者未来对自己造成伤害的可能性(McCullough et al., 2013; 谢东杰, 苏彦捷, 2019)。

第三方惩罚的威慑效应虽然受到普遍认可(e.g., Chen et al., 2014; Fehr & Gächter, 2002; Lieberman & Linke, 2007; Robinson & Darley, 2003; Sell et al., 2009; Yu et al., 2016)，但也有一些研究结果难以用该效应来解释。

首先，根据威慑理论，惩罚力度和惩罚概率共同决定了惩罚的威慑力(Becker, 1968)。因

此，惩罚的力度应该以犯罪者的收益为基准，一般要高于预期收益才能起到威慑作用(Koenig & Riley, 2017)。然而不少研究发现，即使微弱的第三方惩罚也能对规范遵从行为产生积极影响，甚至比严厉惩罚的效果更好(e.g., Kamei, 2018; Tyran & Feld, 2006; Xiao & Houser, 2011; 陈思静 等, 2021)。在这些研究中，对违规者的惩罚力度往往远低于其违规收益，这意味着违背规范仍然是理性经济人的最优选择。如果第三方惩罚仅通过威慑发挥作用，应当无法遏制违背规范行为的发生。然而事实上研究结果却与之相反——即使没有改变潜在违规者的收益结构，第三方惩罚仍然抑制了违规行为。

其次，威慑理论推论人们受到惩罚的概率越高，就越不可能违背规范，然而实证研究结果显示，第三方惩罚的概率与其维护规范的效果之间并没有明确关系。例如，有研究者发现惩罚概率与干预效果之间呈负相关，低概率的第三方惩罚比高概率的惩罚更能提高人们的合作水平(陈思静 等, 2015)；也有研究发现惩罚概率和合作水平之间呈倒 U 形关系，中等概率时效果最好(Qin & Wang, 2013)；还有研究发现第三方惩罚概率对人们的合作水平并没有影响(Stagnaro et al., 2017)。总之，这些与威慑理论假设相悖的研究结果表明，威慑效应可能不是第三方惩罚促使社会成员遵守规范的唯一机制。

显然，威慑理论也无法解释第三方补偿的作用机制。第三方在做出补偿行为时，无论违规者还是旁观者都不会产生惩罚预期，也就不会因害怕惩罚而约束行为。此外，单纯依靠威慑理论也无法解释那些第三方惩罚未能抑制违规行为的研究结果。这意味着，除了威慑效应之外，第三方干预在促进规范遵从方面可能存在其他的作用机制。

4.2 第三方干预的规范信号效应

试想某人因为在公共场合吸烟被他人谴责，他既可能因害怕再次遭到谴责而抑制自己的吸烟行为，也可能会意识到在公共场合吸烟是其他人所反对的行为，多数人都不会这样做，因此自己也不应该这样做。前一种想法暗示了惩罚的威慑效应，后一种想法则意味着人们能够从第三方干预行为中觉察到社会规范的存在，并基于感知规范调整自己的行为。社会规范知觉理论指出，其他群体成员的态度和行为是人们感知社会规范的重要信息来源(Tankard & Paluck, 2016)。尤其是当有人惩罚了违背规范的人，或者恢复了违背规范所造成的消极结果时，能够有效起到重申社会价值、提示社会规范的作用(Peters et al., 2017; Wenzel & Thielmann, 2006)。例如，当人们观察到邻居严格遵守垃圾分类规范时，人们可能隐约意识到将垃圾分类是更可取的行为；而当人们观察到有人因为没有进行垃圾分类被谴责时，人们会更明确地意识到，不进行垃圾分类是错误的。这意味着，第三方干预不仅是一种威胁信号，也是一种规范信号，提示人们当前情境下所隐含的社会规范。

从信息传递的角度来讲，第三方惩罚是反对违背规范行为的明确信号，既表达了惩罚者的道德谴责，也传递了其希望维护的价值观(Kahan, 1996)。惩罚的出现不仅提醒违规者应该及时改变行为，也让其他旁观者意识到这种行为在群体中是不能容忍的，从而强化了社会规范(Xiao & Houser, 2011)。无论惩罚是以社会排斥、流言蜚语还是直接对抗的形式出现，都表达了对违规行为的反对态度，具有澄清、突显社会规范的作用(Eriksson et al., 2021)。这意味着第三方惩罚很可能通过提示社会规范，调整人们对社会规范的知觉，进而改变其行为。

第三方惩罚的规范突显作用得到了一些实证研究的支持。有研究发现，第三方惩罚通过改变人们的规范知觉提高了群体合作水平(陈思静 等, 2021)。相较于第二方惩罚，第三方惩罚是更明确的社会规范信号，观察者可以从中感知到了更积极的社会规范，进而做出更公平的分配(Chen et al., 2020)。也有研究发现，只有公开实施的第三方惩罚才可以促进合作，私下惩罚甚至可能起到反作用(Xiao & Houser, 2011)。这是因为只有公开惩罚才能通过彰显社会规范提高合作水平，而单纯的惩罚威慑无法阻止“搭便车”行为。此外，人们在实施第三方惩罚时也更偏好那些传递了规范信息的惩罚，并且认为那些未能传递规范信息的惩罚方式无法阻止违规行为再次发生(Marshall et al., 2021)。事实也与人们的预期相一致，未提供规范共识的第三方惩罚不仅无法促进合作，反而会加速团队崩溃(Fehr & Williams, 2018)；只有当第三方惩罚和社会规范信息相结合时，才能有效促进人们的合作行为(Andrighetto et al., 2013)、提高互惠水平(Bicchieri et al., 2021)。规范信号视角似乎为第三方惩罚的失效提供了解释：当第三方惩罚出于恶意或过于严厉时，可能会被认为缺乏正当性，导致其澄清社会规范的效果严重削弱。也就是说，当人们否定干预行为的正当性时，就不会将其视为社会规范的信息来源，继而无从感知行为规范要求。显然，在缺乏对恰当行为的认知下，人们很难调整自己的行为以符合社会规范的期望。

与第三方惩罚相类似，第三方补偿作为有代价的信号，同样具有澄清社会价值、突显社会规范的功能(Dhaliwal et al., 2021; Wenzel & Thielmann, 2006)。第三方补偿虽然没有直接谴责违规者，但也隐含了这样的暗示——当前所发生的行为是错误的，所以该行为所造成的后果才需要被弥补和恢复。因此，第三方补偿同样表达了干预者对违背规范行为的反对态度。研究发现，观察到他人捡起地上的垃圾会使得人们关注到命令性社会规范(其他人赞成和不赞成乱扔垃圾的程度)，并且可以有效抑制人们乱扔垃圾的行为(Kallgren et al., 2000; Reno et al., 1993)。该研究表明捡起垃圾作为一种第三方补偿行为，可以成为观察者知觉社会规范的信息来源，传达出行动者对乱扔垃圾的反对态度，进而影响到观察者自身的有关行为。

综上可以推论，规范信号效应似乎也是第三方干预抑制违规行为的重要作用机制。无论

第三方惩罚还是第三方补偿，都可以通过传递社会规范信息改变人们的规范知觉，进而影响其行为。人们可以从惩罚或补偿行为中感知到干预者的心理状态(Gallagher, 2008)，包括干预者对违规行为的反对态度(Nikiforakis & Mitchell, 2014; Reuben & Van Winden, 2008)，以及干预者自己对遵守规范的隐含承诺(Chen et al., 2020; Jordan et al., 2016)，二者分别对应了两种类型的社会规范感知——描述性规范知觉和命令性规范知觉。描述性规范是关于人们实际做什么的规范，描述了特定情境下的典型行为；而命令性规范是关于人们认为应该做什么的规范，描述了特定情境下人们认可或不认可的行为(Cialdini et al., 1990)。大量研究发现，规范知觉对人们的规范遵从行为具有显著的正向作用(e.g., Dieterich et al., 2013; Miller & Prentice, 2016; Paluck et al., 2016; Yitmen & Verkuyten, 2020)。因此，在观察或经历第三方干预后，人们既可以从干预者的隐含承诺中感知到描述性社会规范(即其他社会成员不会做出类似行为)，也可以从干预者的反对态度中感知到命令性社会规范(即其他社会成员认为类似行为是错误的)，进而校正自己的规范知觉并相应地调整自身行为以符合规范要求。反之，如果人们观察到他人对违规行为漠不关心，可能会认为不干预才是当下的群体规范，这将暗示该违规行为是可以接受的，继而反向影响人们对社会规范的感知，最终导致违规行为增加。

5 研究展望

第三方干预在维护社会规范方面所发挥的作用毋庸置疑，以往研究也取得了丰硕的研究成果，然而尚有一些问题未能厘清。首先，当前有关第三方干预维护社会规范的研究大多以惩罚研究为主，对补偿等干预措施影响效应的探讨还相对匮乏。从广义上来讲，任何针对受害者的恢复性措施都属于第三方补偿的范畴，其形式非常多样。在数字化时代下，第三方补偿也衍生出许多低成本、高参与的新形式，如互联网慈善日捐、在线公益课程分享、信息披露与法律援助虚拟社区、社交媒体舆论声援支持等等。由于补偿行为具有纯粹的利他性，也不会如同惩罚一般带来高额附加成本，故而有必要探讨第三方补偿能否成为惩罚的替代性措施，在非正式规范维护系统内发挥同样的作用。比如第三方补偿是否能像第三方惩罚那样，抑制其他社会成员可能的违规行为？如果具有抑制作用，不具威慑力的第三方补偿通过怎样的作用机制施加影响？虽然本文基于一些研究结果进行了理论推演，提出第三方补偿可能作为一种社会规范信号，能够通过改变人们的社会规范知觉来影响其行为，但未来还需要依靠更多实证研究对这一假说加以检验。

除了惩罚与补偿外，奖励可能也是一种重要的第三方干预方式。一项元分析研究发现，与惩罚相类似，奖励同样对社会合作具有中等程度的正向影响作用(Balliet et al., 2011)。如果

说第三方惩罚与补偿均反映了干预者对不良行为的反对态度，那么第三方奖励则反映了干预者对积极行为的肯定与认可。因此第三方奖励不仅是一种外部激励，也可能具有信号作用。未来的研究可以探讨第三方奖励维护社会规范的作用机制，比如奖励是否同样具有传递规范信息、促进规范遵从的功能？以及奖励作为一种外部激励，是否也存在排挤内部动机等潜在的负面效果？

其次，第三方干预影响效应的边界条件还不够清晰。虽然多数研究结果支持了第三方惩罚在促进规范遵从方面的积极作用，但值得注意的是，第三方惩罚并不总是"利他"的，被滥用的惩罚会削弱利他惩罚在维护规范方面的有效性(Pleasant & Barclay, 2018; Rand & Nowak, 2011)。鉴于第三方惩罚可能存在的黑暗面，未来研究不仅有必要寻找惩罚的替代性措施，也有必要探讨第三方惩罚维护社会规范的边界条件。第三方惩罚发生的场景可能会影响人们对惩罚行为道德合法性的判断，如在涉及资源分配的场景中，人们普遍认为应该对不公平的分配者给予惩罚(Martin et al., 2019)；而在涉及合作互动的场景中，人们却并不赞同对未合作者施加惩罚(Sutter et al., 2010)。这意味着，人们并非对所有第三方惩罚行为都盲目认可，而是会对不同场景下惩罚行为的合理性加以判断，并根据情境背景推断惩罚者动机。那么，人们对于第三方惩罚动机与合理性的感知会影响到惩罚效果吗？感知利他的第三方惩罚是否比感知利己的惩罚更好地澄清了社会规范？不同来源及场景下的第三方惩罚是否在传递规范信号的效果上也有所不同？这些问题有待未来研究深入探讨。

此外，第三方干预的成本也可能影响干预效果。一方面，第三方干预的成本越低，干预出现的频率就越高(Guala, 2012)。这意味着在低干预成本时，违规者和受害者得到应得惩罚或帮助的可能性更高，进而干预行为所产生的威慑力会更强、传递规范信号的频率也更高，从而应当在促进规范遵从方面起到更好效果。然而，也有研究发现了不一致的结果：有成本的第三方惩罚比无成本时更能维持和促进合作(Balliet et al., 2011; Kuwabara & Yu, 2017)。这是因为无成本或低成本的惩罚行为可能引发观察者对其道德合法性的质疑（如惩罚者可能出于竞争而非利他做出惩罚），而高成本的惩罚则更能彰显惩罚者的大公无私(Raihani & Bshary, 2015)。后者因此可能被认为更具正当性，并传递出惩罚者无私的、愿意维护社会规范的强烈信号。总之，第三方干预成本对其维护社会规范效果的影响似乎颇为复杂，二者之间的确切关系还有待进一步探究。

再者，第三方干预维护社会规范效果的可持续性还尚不明确。第三方干预毕竟是一种高成本行为，惩罚者可能受到报复或怨恨，补偿者也需要付出时间、精力或金钱成本(Gordon et al., 2014; Wiessner, 2020)。因此，只有当第三方干预能够对人们的有关行为保持长期影响力

时，这种亲社会行为才能在管理实践中更具效率，在社会治理中发挥更大作用。目前还鲜有研究考察了第三方干预的长期影响，虽然有部分研究使用连续多轮博弈任务考察了第三方干预效果的可持续性(e.g., Fehr & Fischbacher, 2004; Gachter et al., 2008; Guerra & Zhuravleva, 2021)，但本质上仍然只检验了第三方干预的短期效果，无法为其长期效力提供证据。未来的研究可以通过多点纵向设计，绘制不同时间距离下人们行为的变化曲线，从而更清晰、准确地了解第三方干预影响的持续时间与效果。

杀鸡儆猴、惩前毖后是为人们所熟知社会互动规则，暗示了惩罚对人们行为的强大影响力。然而，如果人们只是因害怕受到惩罚而选择做一个好人，未免显得人性贫乏、社会僵化。所幸，第三方干预的信号视角为社会规范维系带来了新的契机，初步展现了以非破坏性措施促进社会成员遵从规范的可行性与有效性。通过与数字化大众传媒手段相结合，扩大信号效应的影响范围，第三方干预有望在未来社会治理中发挥“以小拨大”的非凡效力。

参考文献：

陈思静, 朱玥. (2020). 惩罚的另一张面孔:惩罚的负面作用及破坏性惩罚. *心理科学, 43*(4), 911–917.

陈思静, 何铨, 马剑虹. (2015). 第三方惩罚对合作行为的影响：基于社会规范激活的解释. *心理学报, 47*(3), 389–405.

陈思静, 邢懿琳, 翁异静, 黎常. (2021). 第三方惩罚对合作的溢出效应:基于社会规范的解释. *心理学报, 53*(7), 758–772.

陈思静, 徐烨超. (2020). “仁者”还是“智者”：第三方惩罚对惩罚者声誉的影响. *心理学报, 52*(12), 1436–1451.

冯佳兴. (2020). 第三方利他行为的心理机制——基于社会规范的分析. *心理学进展, 10*(3), 252–259.

刘映杰, 段亚妮, 刘昊馨, 刘佳, 王赫. (2022). 得失情境下第三方惩罚决策差异的神经机制:基于 rtms 的研究. *心理科学, 45*(4), 942–952.

王华根, 甄珍, 刘超, 秦绍正. (2020). 锄强还是扶弱:急性应激如何影响第三方决策. *科学通报, 65*(19), 1975–1984.

谢东杰, 苏彦捷. (2019). 第三方惩罚的演化与认知机制. *心理科学, 42*(1), 216–222.

徐杰, 孙向超, 董悦, 汪祚军, 李伟强, 袁博. (2017). 人情与公正的抉择:社会距离对第三方干预的影响. *心理科学, 40*(5), 1175–1181.

Akers, R. L. (1990). Rational choice, deterrence, and social learning theory in criminology: The path not taken. *The Journal of Law and Criminology, 81*(3), 653–676.

Almenberg, J., Dreber, A., Apicella, C.L., & Rand, D.G. (2010). *Third party reward and punishment: Group size, efficiency and public goods*. CSN: Economics (Topic).

Andrighetto, G., Brandts, J., Conte, R., Sabater-Mir, J., Solaz, H., & Villatoro, D. (2013). Punish and voice: punishment enhances cooperation when combined with norm-signalling. *PloS One, 8*(6), e64941.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications, 7*(1), 1–6.

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137*(4), 594–615.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Basurto, X., Blanco, E., Nenadovic, M., & Vollan, B. (2016). Integrating simultaneous prosocial and antisocial behavior into theories of collective action. *Science Advances, 2*(3), e1501220.

Baumgartner, T., Götte, L., Gügler, R., & Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Human Brain Mapping, 33*(6), 1452–1469.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy, 76*(2), 169–217.

Bellucci, G., Camilleri, J. A., Iyengar, V., Eickhoff, S. B., & Krueger, F. (2020). The emerging neuroscience of social punishment: Meta-analytic evidence. *Neuroscience & Biobehavioral Reviews, 113*, 426–439.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature, 442*(7105), 912–915.

Bicchieri, C., Dimant, E., & Xiao, E. (2021). Deviant or wrong? The effects of norm information on the efficacy of punishment. *Journal of Economic Behavior & Organization, 188*, 209–235.

Bottom, W. P., Gibson, K., Daniels, S. E., & Murnighan, J. K. (2002). When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation. *Organization Science, 13*(5), 497–513.

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology, 42*(4), 437–451.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83*(2), 284–299.

Charness, G., Cobo-Reyes, R., & Jiménez, N. (2008). An investment game with third-party intervention. *Journal of Economic Behavior & Organization, 68*(1), 18–28.

Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology, 39*, 268–277.

Chen, H., Zeng, Z., & Ma, J. (2020). The source of punishment matters: Third-party punishment restrains observers from selfish behaviors better than does second-party punishment by shaping norm perceptions. *PloS One, 15*(3), e0229510.

Chen, X., Szolnoki, A., & Perc, M. (2014). Probabilistic sharing solves the problem of costly punishment. *New Journal of Physics, 16*(8), 083016.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015–1026.

De Cremer, D. (2010). To pay or to apologize? On the psychology of dealing with unfair offers in a dictator game. *Journal of Economic Psychology, 31*(6), 843–848.

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters

for third-party punishment. *Evolution and Human Behavior, 38*(6), 734–743.

Dhaliwal, N. A., Patil, I., & Cushman, F. (2021). Reputational and cooperative benefits of third-party compensation. *Organizational Behavior and Human Decision Processes, 164*, 27–51.

Dieterich, S. E., Stanley, L. R., Swaim, R. C., & Beauvais, F. (2013). Outcome expectancies, descriptive norms, and alcohol use: American Indian and white adolescents. *The Journal of Primary Prevention, 34*(4), 209–219.

Dimitroff, S. J., Harrod, E. G., Smith, K. E., Faig, K. E., Decety, J., & Norman, G. J. (2020). Third-party punishment following observed social rejection. *Emotion, 20*(4), 713.

Eriksson, K., Strimling, P., Gelfand, M., Wu, J., Abernathy, J., Akotia, C. S., ... & Van Lange, P. A. (2021). Perceptions of the appropriate response to norm violation in 57 societies. *Nature Communications, 12*(1), 1–11.

Fabbri, M., & Carbonara, E. (2017). Social influence on third-party punishment: An experiment. *Journal of Economic Psychology, 62*, 204–230.

Fatas, E., & Mateu, G. (2015). Antisocial punishment in two social dilemmas. *Frontiers in Behavioral Neuroscience, 9*, 107.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63–87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137–140.

Fehr, E., & Williams, T. (2018). *Social norms, Endogenous sorting and the culture of cooperation*. (ECON Working Papers 267). Zurich, Switzerland: Department of Economics of University of Zurich.

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science, 25*(3), 656–664.

Gachter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science, 322*, 1510–1510.

Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition, 17*(2), 535–543.

Gibbs, J. P. (1968). Crime, punishment, and deterrence. *The Southwestern Social Science Quarterly*, 515–530.

Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., & Marois, R. (2016). Parsing the behavioral and brain mechanisms of third-party punishment. *Journal of Neuroscience, 36*(36), 9420–9434.

Gordon, D. S., Madden, J. R., & Lea, S. E. (2014). Both loved and feared: Third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. *PloS One,*

9(10), e110045.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and brain sciences, 35*(1), 1–15.

Guerra, A., & Zhuravleva, T. (2021). Do bystanders react to bribery?. *Journal of Economic Behavior & Organization, 185*, 442–462.

Gummerum, M., Van Dillen, L. F., Van Dijk, E., & López-Pérez, B. (2016). Costly third–party interventions: The role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *Journal of Experimental Social Psychology, 65*, 94–104.

Hart, E., Mellers, B. A., & Bicchieri, C. (2019). Bad luck or bad intentions: When do third parties reveal offenders' intentions to victims?. *Journal of Experimental Social Psychology, 84*, 103788.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: The emergence of costly punishment. *Science, 316*, 1905–1907.

Heffner, J., & FeldmanHall, O. (2019). Why we don't always punish: Preferences for non-punitive responses to moral violations. *Scientific Reports, 9*(1), 1–13.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... & Ziker, J. (2006). Costly punishment across human societies. *Science, 312*, 1767–1770.

Herrmann, B., Thoni, C., & Gachter, S. (2008). Antisocial punishment across societies. *Science, 319*, 1362–1367.

Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: Neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience, 9*, 24.

Hume, D. (1888). *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects*, ed. LA Selby–Bigge. New York: Oxford Univ. Press.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*, 473–476.

Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences, 111*(35), 12710–12715.

Kahan, D. M. (1996). What Do Alternative Sanctions Mean?. *The University of Chicago Law Review, 63*(2), 591–653.

Kallgren, C. A., Reno, R. R., & Cialdini, R. B. (2000). A focus theory of normative conduct: When norms do and do not affect behavior. *Personality and Social Psychology Bulletin, 26*(8), 1002–1012.

Kamei, K. (2018). The role of visibility on third party punishment actions for the enforcement of social norms.

Economics Letters, 171, 193–197.

Kamei, K. (2020). Group size effect and over-punishment in the case of third party enforcement of social norms. *Journal of Economic Behavior & Organization, 175*, 395–412.

Koenig, B. L., & Riley, C. M. (2017). To what reference point do people calibrate cost-free, third-party punishment?. *Personality and Individual Differences, 115,* 90–98.

Kuwabara, K., & Yu, S. (2017). Costly punishment increases prosocial punishment by designated punishers: Power and legitimacy in public goods games. *Social Psychology Quarterly, 80*(2), 174–193.

Leliveld, M. C., van Dijk, E., & van Beest, I. (2012). Punishing and compensating others at your own expense: The role of empathic concern on reactions to distributive injustice. *European Journal of Social Psychology, 42*(2), 135–140.

Lergetporer, P., Angerer, S., Glätzle-Rützler, D., & Sutter, M. (2014). Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proceedings of the National Academy of Sciences, 111*(19), 6916–6921.

Lewisch, P. G., Ottone, S., & Ponzano, F. (2011). Free-riding on altruistic punishment? An experimental comparison of third-party punishment in a stand-alone and in an in-group environment. *Review of Law & Economics, 7*(1), 161–190.

Lieberman, D., & Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology, 5*(2), 289–305.

Liu, Y., Bian, X., Hu, Y., Chen, Y. T., Li, X., & Di Fabrizio, B. (2018). Intergroup bias influences third-party punishment and compensation: In-group relationships attenuate altruistic punishment. *Social Behavior and Personality: An International Journal, 46*(8), 1397–1408.

Liu, Y., Wang, H., Li, L., Wang, Y., Peng, J., & Baxter, D. F. (2019). Judgments in a hurry: Time pressure affects how judges assess unfairly shared losses and unfairly shared gains. *Scandinavian Journal of Psychology, 60*(3), 203–212.

Lotz, S., Okimoto, T. G., Schlösser, T., & Fetchenhauer, D. (2011). Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *Journal of Experimental Social Psychology, 47*(2), 477–480.

Malouff, J., Thorsteinsson, E., Schutte, N., & Rooke, S. E. (2009). Effects of vicarious punishment: A meta-analysis. *The Journal of General Psychology, 136*(3), 271–286.

Marshall, J., Gollwitzer, A., Wynn, K., & Bloom, P. (2019). The development of corporal third-party punishment.

Cognition, 190, 221–229.

Marshall, J., Yudkin, D. A., & Crockett, M. J. (2021). Children punish third parties to satisfy both consequentialist and retributive motives. *Nature Human Behaviour, 5*(3), 361–368.

Martin, J. W., Jordan, J. J., Rand, D. G., & Cushman, F. (2019). When do we punish people who don't. *Cognition, 193*, 104040.

Martin, J. W., Martin, S., & McAuliffe, K. (2021). Third-party punishment promotes fairness in children. *Developmental Psychology, 57*(6), 927–939.

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences, 108*(28), 11375–11380.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences, 36*(1), 1–15.

Miller, D. T., & Prentice, D. A. (2016). Changing norms to change behavior. *Annual Review of Psychology, 67*, 339–361.

Molho, C., Tybur, J. M., Van Lange, P. A., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature Communications, 11*(1), 1–9.

Mulder, L. B., Van Dijk, E., De Cremer, D., & Wilke, H. A. (2006). Undermining trust and cooperation: The paradox of sanctioning systems in social dilemmas. *Journal of Experimental Social Psychology, 42*(2), 147–162.

Nikiforakis, N., & Mitchell, H. (2014). Mixing the carrots with the sticks: Third party punishment and reward. *Experimental Economics, 17*(1), 1–23.

Ouyang, H., Yu, J., Duan, J., Zheng, L., Li, L., & Guo, X. (2021). Empathy-based tolerance towards poor norm violators in third-party punishment. *Experimental Brain Research, 239*(7), 2171–2180.

Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences, 113*(3), 566–571.

Peters, K., Jetten, J., Radova, D., & Austin, K. (2017). Gossiping about deviance: Evidence that deviance spurs the gossip that builds bonds. *Psychological Science, 28*(11), 1610–1619.

Pleasant, A., & Barclay, P. (2018). Why hate the good guy? Antisocial punishment of high cooperators is greater when people compete to be chosen. *Psychological Science, 29*(6), 868–876.

Qin, X., & Wang, S. (2013). Using an exogenous mechanism to examine efficient probabilistic punishment. *Journal of Economic Psychology, 39*, 1–10.

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in ecology & evolution, 30*(2), 98–103.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science, 325*, 1272–1275.

Rand, D. G., & Nowak, M. A. (2011). The evolution of antisocial punishment in optional public goods games. *Nature Communications, 2*(1), 434.

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology, 64*(1), 104–112.

Reuben, E., & Van Winden, F. (2008). Social ties and coordination on negative reciprocity: The role of affect. *Journal of Public Economics, 92*(1–2), 34–53.

Robinson, P. H., & Darley, J. M. (2003). The role of deterrence in the formulation of criminal law rules: At its worst when doing its best. *The Georgetown Law Journal, 91*, 949–1002.

Rodrigues, J., Liesner, M., Reutter, M., Mussel, P., & Hewig, J. (2020). It's costly punishment, not altruistic: Low midfrontal theta and state anger predict punishment. *Psychophysiology, 57*(8), e13557.

Schroeder, D. A., Steel, J. E., Woodell, A. J., & Bembenek, A. F. (2003). Justice within social dilemmas. *Personality and Social Psychology Review, 7*(4), 374–387.

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences, 106*(35), 15073–15078.

Stagnaro, M. N., Arechar, A. A., & Rand, D. G. (2017). From good institutions to generous citizens: Top-down incentives to cooperate promote subsequent prosociality but not norm enforcement. *Cognition, 167*, 212–254.

Sutter, M., Haigner, S., & Kocher, M. G. (2010). Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *The Review of Economic Studies, 77*(4), 1540–1566.

Sylwester, K., Herrmann, B., & Bryson, J. J. (2013). Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics, 6*(3), 167–188.

Tan, F., & Xiao, E. (2018). Third-party punishment: Retribution or deterrence?. *Journal of Economic Psychology, 67*, 34–46.

Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review, 10*(1), 181–211.

Thulin, E. W., & Bicchieri, C. (2016). I'm so angry I could help you: Moral outrage as a driver of victim compensation. *Social Philosophy and Policy, 32*(2), 146–160.

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology, 64*,

231–255.

Tyran, J. R., & Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *The Scandinavian Journal of Economics, 108*(1), 135–156.

van Dijk, E., Molenmaker, W. E., & de Kwaadsteniet, E. W. (2015). Promoting cooperation in social dilemmas: The use of sanctions. *Current Opinion in Psychology, 6*, 118–122.

Van Doorn, J., Zeelenberg, M., & Breugelmans, S. M. (2014). Anger and prosocial behavior. *Emotion Review, 6*(3), 261–268.

Van Doorn, J., Zeelenberg, M., Breugelmans, S. M., Berger, S., & Okimoto, T. G. (2018). Prosocial consequences of third-party anger. *Theory and Decision, 84*(4), 585–599.

van Prooijen, J. W. (2010). Retributive versus compensatory justice: Observers' preference for punishing in response to criminal offenses. *European Journal of Social Psychology, 40*(1), 72–85.

Wenzel, M., & Thielmann, I. (2006). Why we punish in the name of justice: Just desert versus value restoration and the role of social identity. *Social Justice Research, 19*(4), 450–470.

Wiessner, P. (2020). The role of third parties in norm enforcement in customary courts among the Enga of Papua New Guinea. *Proceedings of the National Academy of Sciences, 117*(51), 32320–32328.

Xiao, E., & Houser, D. (2011). Punish in public. *Journal of Public Economics, 95*(7–8), 1006–1017.

Yitmen, Ş., & Verkuyten, M. (2020). Support to Syrian refugees in Turkey: The roles of descriptive and injunctive norms, threat, and negative emotions. *Asian Journal of Social Psychology, 23*(3), 293–301.

Yu, T., Chen, S. H., & Li, H. (2016). Social norms, costly punishment and the evolution of cooperation. *Journal of Economic Interaction and Coordination, 11*(2), 313–343.

Deterrence or signal? The function of third-party intervention

GUO Yuchen, LIU Yanbin, CHENG Yuan

School of Business, NingboTech University, Ningbo 315100, China

Abstract: Third-party intervention is crucial in maintaining and developing social norms with evolutionary implications for humans. Compensation and punishment are the two main forms of third-party intervention and important forces for maintaining social norms by restoring the balance of gains and losses and promoting norm compliance. The deterrent effect is widely recognized by researchers as the main mechanism by which third-party punishment promotes norm compliance. However, several studies contradict this hypothesis. Moreover, as a costly signal, third-party interventions can clarify social norms and adjust individuals' perceptions of them. This indicates that the signaling effect may also be an important mechanism for third-party interventions to promote norm compliance. Future research must explore the boundary conditions for the impact of third-party punishment on norm compliance and test the effectiveness of third-party compensation in maintaining social norms.

Key words: third-party intervention, third-party punishment, third-party compensation, social norms, deterrence effect, signaling effect